Enhancing Dialectal Arabic Intent Detection through Cross-Dialect Multilingual Input Augmentation

Shehenaz Hossain[1], Fouad Shammary[2], Bahaulddin Shammary[1], Haithem Afli[1],

[1]ADAPT Centre, Munster Technological University, Cork, Ireland
[2]Alef Education, Abu Dhabi, United Arab Emirates

Correspondence: shehenaz.hossain@mycit.ie, fouad.shammary@alefeducation.com, bahaulddin.shammary@mycit.ie, Haithem.Afli@mtu.ie

Abstract

Addressing the challenges of Arabic intent detection amid extensive dialectal variation, this study presents a crossdialtectal, multilingual approach for classifying intents in banking and migration contexts. By augmenting dialectal inputs with Modern Standard Arabic (MSA) and English translations, our method leverages cross-lingual context to improve classification accuracy. We evaluate single-input (dialect-only), dual-input (dialect + MSA), and triple-input (dialect + MSA + English) models, applying language-specific tokenization for each. Results demonstrate that, in the migration dataset, our model achieved an accuracy gain of over 50% on Tunisian dialect, increasing from 43.3% with dialect-only input to 94% with the full multilingual setup. Similarly, in the PAL (Palestinian dialect) dataset, accuracy improved from 87.7% to 93.5% with translation augmentation, reflecting a gain of 5.8 percentage points. These findings underscore the effectiveness of our approach for intent detection across various Arabic dialects.

1 Introduction

Natural language understanding (NLU) powers the smart applications we use daily by helping machines grasp human intent. Yet, intent detection in Arabic is especially challenging due to the language's diversity. Spoken by over 400 million people across 22+ countries, Arabic includes both Modern Standard Arabic (MSA) for formal use and a variety of regional dialects (Al' Ammiya) for everyday speech. Each dialect presents unique challenges, from vocabulary and grammar to pronunciation, making intent detection a complex task.

This linguistic diversity poses a significant challenge for NLU systems. For example, the phrase "illegal migration" translates to الهجرة غير الشرعية (al-hijra ghayr al-shar'iyya) in MSA, but in Moroccan Arabic, it's الحريگ (al-harq), while in Tunisian Arabic, it's الحرقة (al-harqa). Without standardized spelling or structure across dialects, NLP models face a daunting task, as spelling inconsistencies and borrowed words from other languages often add extra layers of complexity.

This study introduces a novel approach that combines multilingual and multidialectal strategies to detect intent in banking (PAL dataset) and migration contexts (GPT-generated Migration dataset). Each dialectal text is translated into MSA and English to see if the structural clarity of MSA and the broader context of English enhance intent recognition. MSA adds consistency, while English captures additional meaning that might otherwise be missed.

2 Related Work

Multidialectal intent detection in Arabic presents unique challenges due to the diverse dialects and limited annotated data. Early research predominantly focused on MSA, but the advent of pretrained models like BERT opened new avenues for Arabic NLP. Francony et al. 2019 addressed the issue of dialect diversity by proposing a hierarchical deep learning framework. Their two-step model distinguishes MSA from dialects and further classifies dialects by region, offering a structured foundation for tasks like intent detection. Similarly, Shammary et al. 2022 explored a comparative analysis of traditional TF-IDF approaches and transformer-based models for the NADI 2022 shared task. Their findings demonstrated that while transformers are powerful, TF-IDF can be a competitive and lightweight alternative for low-resource dialects, emphasizing the value of efficient methods in resource-constrained settings. Al Hariri and Abu Farha 2024 showed that Arabic BERT models, while effective for MSA, required additional fine-tuning for dialectal Arabic. To address this, Elkordi et al. 2024 introduced contrastive learning techniques to detect intents

Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4), pages 44–49
January 20th, 2025.

in different dialects, particularly in the banking sector, while Ramadan et al. 2024 developed a BERT-based ensemble model for the detection of cross-dialectal intent. One major challenge is data scarcity. Duwairi and Abushaqra 2021 addressed this issue through back-translation and paraphrasing, improving performance for low-resource dialects like Moroccan and Sudanese Arabic. Similarly, El-Makky et al. 2024 explored transfer learning to fine-tune models trained on high-resource dialects and apply them to others, enhancing generalization across dialects. To further address dialect-specific challenges, Skiredj et al. 2024 introduced the DarijaBanking dataset for Moroccan Arabic intent detection in the banking domain. The study presents BERTouch, a Darija-specific BERT model achieving state-of-the-art performance. Their findings highlight the need for domain-specific resources and multilingual approaches for effective intent detection. Shared tasks like AraFinNLP Malaysha et al. 2024a have provided benchmark datasets for multidialectal Arabic NLP. These challenges have helped researchers explore advanced techniques such as pre-trained models and data augmentation to enhance performance in intent detection for Arabic dialects. Fares and Touileb 2024 fine-tuned a T5 model and generated synthetic data in Moroccan, Tunisian, and Saudi dialects. By leveraging model ensembling, they highlighted synthetic data's role in handling dialectal variation.

3 Dataset

Our study draws on two primary datasets: The first is ArBanking77 (Jarrar et al. 2023) provided for the shared task 1 of the AraFinNLP 2024 (Malaysha et al. 2024b), which contains Arabic banking queries in both MSA and the Palestinian (PAL) dialect, labeled across 77 intent categories. Our analysis used only the PAL subset, which focuses on the Palestinian dialect. The second dataset was generated [1] with GPT-4, centering on Tunisian dialect text related to illegal migration. This dataset is labeled by intent strength—categorized as non-intent, weak intent, or strong intent—and each entry was meticulously validated by a native Tunisian speaker with specialized expertise in dialectal nuances and migration-related terminology, ensuring both linguistic fidelity and contextual depth.

[1]The dataset created for this research will be publicly available upon publication.

Table 1 shows the original ArBanking77 Dataset distribution.

Dialect	Train	Dev
MSA	10733	1230
PAL	10821	1234

Table 1: Dataset Statistics of ArBanking77

For our experiments, we split the PAL training set into 85% for training and 15% for testing, using the dev set for validation. The migration dataset was divided into 70% training, 15% validation, and 15% testing, ensuring balanced evaluation across intent strength labels.

Table 2 Shows the sample distribution in both of the datasets.

Dataset	Total	Train	Val	Test
PAL	10821	9197	1234	1624
Migration	2000	1398	300	300

Table 2: Dataset Distribution

4 Methodology

In this section, we detail our methodology for developing models capable of detecting intents in multi-dialectal banking and migration datasets. Our approach combines translation, tokenization, and model configurations designed to harness the benefits of Modern Standard Arabic (MSA) and English alongside dialectal inputs.

4.1 Translation

For the translation component, we utilized two open-source models via Hugging Face: Murhaf/AraT5-MSAizer[2] (Fares 2024) for Arabic dialect-to-MSA translation and Helsinki-NLP/opus-mt-ar-en[3] (Tiedemann and Thottingal 2020) for Arabic dialect-to-English translation. Both models are freely accessible on the Hugging Face platform, streamlining their integration into our workflow. AraT5-MSAizer[4], a fine-tuned version of UBC-NLP/AraT5v2-base-1024, is optimized for regional Arabic dialects (e.g., Levantine, Maghrebi, Gulf) and achieved a BLEU score of 21.79 on the OSACT 2024 test set, indicating reliable MSA translations that clarify dialectal

[2]AraT5-MSAizer on Hugging Face
[3]Helsinki-NLP/opus-mt-ar-en on Hugging Face
[4]Github.com/AraT5-MSAizer

ambiguities. Meanwhile, Helsinki-NLP/opus-mt-ar-en, part of the Opus-MT project[5] (Tiedemann 2020), is highly effective for Arabic-to-English translation, achieving a BLEU score of 49.4 on the Tatoeba test set. While primarily trained on MSA, it leverages multilingual data that may include elements of dialectal Arabic, making it useful for capturing semantic nuances in dialects. Its open-source nature and ease of deployment make it highly practical for resource-constrained settings.

4.2 Tokenization

To ensure consistency across data sources prior to tokenization, we pre-processed both datasets. In the PAL dataset, intents were mapped to integers(0-76), for 77 financial service-related categories, with missing or invalid entries removed, with the missing or invalid entries removed. For the Migration dataset, the intentions were categorized by strength: Non-intention (0), weak intention (1) and strong intention (2), and invalid entries were excluded. We applied specialized tokenizers to each language variant to capture the unique linguistic nuances of Arabic dialects, MSA, and English, accommodating the significant divergence of dialectal Arabic from MSA. For dialectal Arabic, we used the CAMeLBERT-Mix (bert-base-arabic-camelbert-mix)[6](Inoue et al. 2021)tokenizer derived from CAMeLBert-mix model, pretrained on a mixture of Arabic texts with different sizes and variants like MSA, DA, and classical Arabic. For MSA texts, we used MARBERT[7](Abdul-Mageed, Elmadany, and Nagoudi 2021)tokenizer derived from MARBERT, a model specifically trained on MSA and DA and proficient in capturing formal Arabic semantics. MARBERT's MSA-focused vocabulary and embeddings allowed us to standardize the input content, providing a consistent Arabic representation across both datasets. This step was particularly useful for understanding how formalized language influences intent classification in contrast to the colloquial forms in dialect. To process the English translations, we used the BERT-base model(uncased)[8](Devlin et al. 2018) tokenizer derived from BERT, a widely adopted English language model capable of extracting semantic information from English text. All inputs were tokenized with a maximum sequence length of 128 tokens using padding and truncation for consistency across input sizes.

5 Model Architecture

We developed three configurations: the Dialect-Only Model (DOM), the Dialect-MSA Model (DMM) (dialect + MSA), and the Dialect-MSA-English Model (DMEM) (dialect + MSA + English). These configurations allow us to assess whether adding MSA and English translations enhances model performance.. Figure1 depicts an outline of our approach.

5.1 Dialect-Only Model (DOM)

This configuration uses only the original dialect input, encoded by CAMeLBERT-Mix (bert-base-arabic-camelbert-mix) for DA. The `[CLS]`[9]token (768 dimensions) is fed into a fully connected layer for classification, with dropout rates of 0.3 for the Migration dataset and 0.1 for PAL to reduce overfitting. The model outputs logits for 3 intent classes in Migration and 77 in PAL.

5.2 Dialect-MSA Model (DMM)

In this configuration, we combine dialect input with its MSA translation, encoded by CAMeLBERT-Mix and MARBERT respectively. The `[CLS]` tokens (768 dimensions each) are concatenated into a 1536-dimensional vector, and fed into a fully connected layer for classification. Like the DOM, dropout rates of 0.3 for Migration and 0.1 for PAL are applied to reduce overfitting, allowing the model to leverage both dialectal and formal Arabic.

5.3 Dialect-MSA-English Model (DMEM)

This model extends the previous configurations(5.1 and 5.2) by incorporating the original dialect input, its MSA translation, and an English translation. CAMeLBERT-Mix encodes the dialectal Arabic, MARBERT encodes MSA, and BERT-base-uncased processes the English translation. The `[CLS]` tokens from each encoder (768 dimensions each) are concatenated into a 2304-dimensional vector, passed through a fully connected layer for classification. As with DOM and

[5] github.com/Helsinki-NLP/Opus-MT

[6] CAMeL-Lab/bert-base-arabic-camelbert-mix on Hugging Face

[7] MARBERT on Hugging Face

[8] BERT-base-uncased on Hugging Face

[9] CLS tokens are special tokens placed at a beginning of each input example in a BERT model, providing a representation of the entire input for use in classification tasks(Devlin et al. 2018).

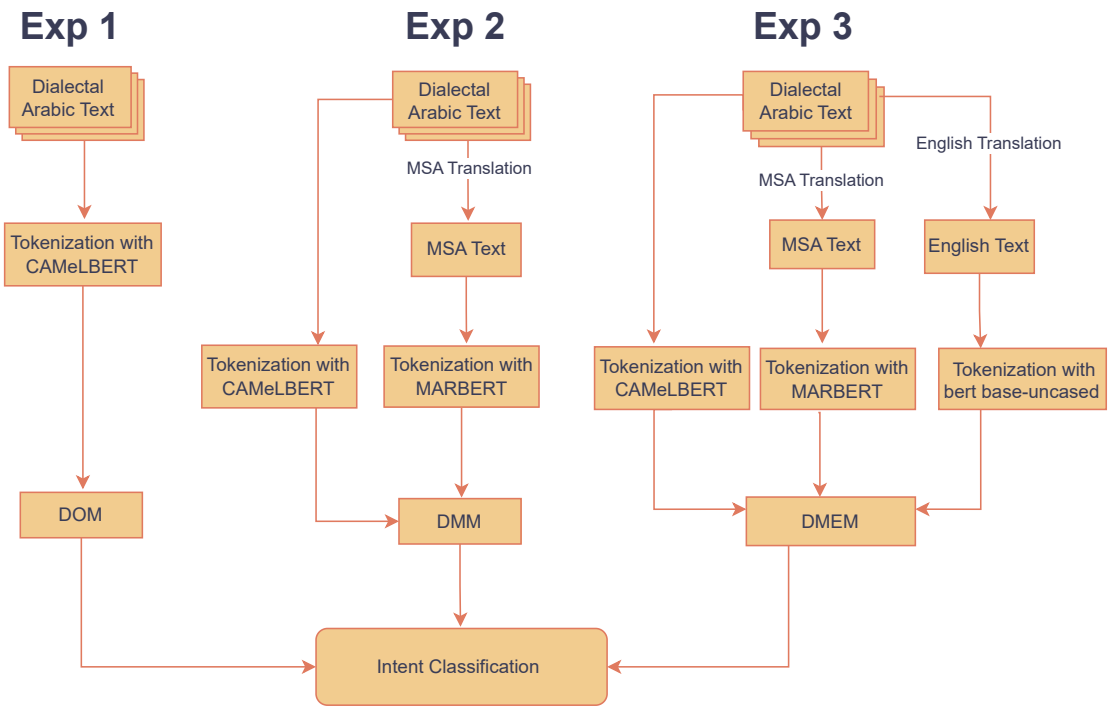

Figure 1: An outline of our approach

DMM, dropout rates are set at 0.3 for Migration and 0.1 for PAL to reduce overfitting. This multilingual configuration enables the model to leverage dialectal, formal Arabic, and English contexts, further enhancing intent classification accuracy.

6 Experimental Setup

We trained models on both datasets (PAL and Migration) with a batch size of 8, using the AdamW optimizer. The learning rate was set to 1e-5 for the Migration and 3e-5 for PAL, with a weight decay of 1e-4 for regularization. Cross-entropy loss was used for multiclass classification. To improve generalization with limited data in the Migration dataset (2000 samples), we applied layer freezing to the lower layers of the CAMeLBERT, MARBERT, and BERT encoders, preserving their pretrained linguistic embeddings and focusing optimization on the task-specific upper layers. Gradient clipping with a maximum norm of 1.0 was implemented to stabilize training, which is especially beneficial in a multi-encoder setup. Early stopping with a patience of 2 epochs was applied, and all models were trained for 3 epochs to balance performance and computational efficiency.

7 Results and Discusssion

Table 3 and Table 4 present the performance metrics across different model configurations for the Migration and PAL datasets.

Config	Acc	Prec	Rec	F1
DOM	0.433	0.70	0.43	0.35
DMM	0.843	0.87	0.84	0.85
DMEM	0.940	0.94	0.94	0.94

Table 3: Performance Metrics (Macro Average) for the Migration Dataset using DOM, DMM, and DMEM.

Config	Acc	Prec	Rec	F1
DOM	0.877	0.88	0.87	0.87
DMM	0.893	0.89	0.89	0.89
DMEM	0.935	0.93	0.92	0.92

Table 4: Performance Metrics (Macro Average) for the PAL Dataset using DOM, DMM, and DMEM.

This study aimed to enhance intent classification across Arabic dialects by incorporating MSA and English translations alongside dialectal Arabic inputs. Results show that the Dialect-Only Model (DOM) provides a baseline with moderate performance (43.3% accuracy for Migration and 87.7% for PAL). Adding MSA translations in the

Dialect-MSA Model (DMM) raised accuracy to 84.3% for Migration and 89.3% for PAL, indicating that the formal structure of MSA helps clarify dialectal ambiguities. Introducing English translations in the Dialect-MSA-English Model (DMEM) further increased accuracy to 94.0% for Migration and 93.5% for PAL, where the cross-lingual context aids with domain-specific terminology in finance and migration.

While the results are promising, limitations emerge due to the MSA bias of pre-trained models like CAMeLBERT and MARBERT. These models, though trained on a mix of Arabic dialects and MSA, still favour MSA, posing challenges, particularly with the Tunisian dialect, which is underrepresented in training data. The distinct vocabulary, syntax, and colloquial phrases of Tunisian diverge significantly from MSA and other Arabic dialects, causing occasional misclassifications and reducing interpretability on migration-related topics. These findings suggest that fine-tuning models on underrepresented dialects, such as Tunisian, may improve intent classification in dialect-heavy datasets, especially those with high linguistic variability.

8 Conclusion and Future Work

Our findings show that adding MSA and English translations to dialectal Arabic improves intent classification. However, challenges persist due to the MSA bias in pre-trained models, impacting performance, particularly for the Tunisian dialect. Expanding training to cover more dialects could help create a more inclusive model. Additionally, fine-tuning large language models on dialectal Arabic holds promise. This approach may better capture linguistic and cultural nuances, enabling more accurate and adaptable intent classification across diverse Arabic-speaking communities.

Acknowledgments

Data generation for the Tunisian dialect and code development for this work were assisted by GPT-3.5 Turbo and GPT-4 Omni. This research is supported by Taighde Éireann – Research Ireland through ADAPT Centre (Grant No. 13/RC/2106_P2) (www.adaptcentre.ie) at Munster Technological University.

References

Abdul-Mageed, Muhammad, AbdelRahim Elmadany, and El Moatez Billah Nagoudi (Aug. 2021). "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 7088–7105. DOI: 10.18653/v1/2021.acl-long.551. URL: https://aclanthology.org/2021.acl-long.551.

Al Hariri, Youssef and Ibrahim Abu Farha (Aug. 2024). "SMASH at AraFinNLP2024: Benchmarking Arabic BERT models on the intent detection." English. In: *Proceedings of The Second Arabic Natural Language Processing Conference*. The Second Arabic Natural Language Processing Conference, ArabicNLP 2024 ; Conference date: 16-08-2024 Through 16-08-2024. Association for Computational Linguistics (ACL), pp. 403–409. URL: https://arabicnlp2024.sigarab.org/.

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

Duwairi, Rana and Feras Abushaqra (2021). "Syntactic- and morphology-based text augmentation framework for Arabic sentiment analysis." In: *PeerJ Computer Science* 7, e469. DOI: 10.7717/peerj-cs.469. URL: https://doi.org/10.7717/peerj-cs.469.

Elkordi, Hossam et al. (Aug. 2024). "AlexuNLP24 at AraFinNLP2024: Multi-Dialect Arabic Intent Detection with Contrastive Learning in Banking Domain." In: *Proceedings of The Second Arabic Natural Language Processing Conference*. Ed. by Nizar Habash et al. Bangkok, Thailand: Association for Computational Linguistics, pp. 415–421. URL: https://aclanthology.org/2024.arabicnlp-1.37.

Fares, Murhaf (May 2024). "AraT5-MSAizer: Translating Dialectal Arabic to MSA." In: *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @*

LREC-COLING 2024. Ed. by Hend Al-Khalifa et al. Torino, Italia: ELRA and ICCL, pp. 124–129. URL: https://aclanthology.org/2024.osact-1.16.

Fares, Murhaf and Samia Touileb (Aug. 2024). "BabelBot at AraFinNLP2024: Fine-tuning T5 for Multi-dialect Intent Detection with Synthetic Data and Model Ensembling." In: *Proceedings of The Second Arabic Natural Language Processing Conference*. Ed. by Nizar Habash et al. Bangkok, Thailand: Association for Computational Linguistics, pp. 433–440. DOI: 10.18653/v1/2024.arabicnlp-1.40. URL: https://aclanthology.org/2024.arabicnlp-1.40.

Francony, Gael de et al. (Aug. 2019). "Hierarchical Deep Learning for Arabic Dialect Identification." In: *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. Ed. by Wassim El-Hajj et al. Florence, Italy: Association for Computational Linguistics, pp. 249–253. DOI: 10.18653/v1/W19-4631. URL: https://aclanthology.org/W19-4631.

Inoue, Go et al. (Apr. 2021). "The Interplay of Variant, Size, and Task Type in Arabic Pretrained Language Models." In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Online): Association for Computational Linguistics.

Jarrar, Mustafa et al. (Dec. 2023). "ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic." In: *Proceedings of ArabicNLP 2023*. Ed. by Hassan Sawaf et al. Singapore (Hybrid): Association for Computational Linguistics, pp. 276–287. DOI: 10.18653/v1/2023.arabicnlp-1.22. URL: https://aclanthology.org/2023.arabicnlp-1.22.

El-Makky, Ahmed et al. (2024). "Transfer Learning for Dialect Generalization: Fine-Tuning Models on High-Resource Dialects." In: *Journal of Computational Linguistics* 50.2, pp. 123–145.

Malaysha, Sanad et al. (2024a). *AraFinNLP 2024: The First Arabic Financial NLP Shared Task*. arXiv: 2407.09818 [cs.CL]. URL: https://arxiv.org/abs/2407.09818.

– (Aug. 2024b). "AraFinNLP 2024: The First Arabic Financial NLP Shared Task." In: *Proceedings of The Second Arabic Natural Language Processing Conference*. Ed. by Nizar Habash et al. Bangkok, Thailand: Association for Computational Linguistics, pp. 393–402. DOI: 10.18653/v1/2024.arabicnlp-1.34. URL: https://aclanthology.org/2024.arabicnlp-1.34.

Ramadan, Asmaa et al. (Aug. 2024). "MA at AraFinNLP2024: BERT-based Ensemble for Cross-dialectal Arabic Intent Detection." In: *Proceedings of The Second Arabic Natural Language Processing Conference*. Ed. by Nizar Habash et al. Bangkok, Thailand: Association for Computational Linguistics, pp. 441–445. URL: https://aclanthology.org/2024.arabicnlp-1.41.

Shammary, Fouad et al. (Dec. 2022). "TF-IDF or Transformers for Arabic Dialect Identification? ITFLOWS participation in the NADI 2022 Shared Task." In: *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*. Ed. by Houda Bouamor et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 420–424. DOI: 10.18653/v1/2022.wanlp-1.42. URL: https://aclanthology.org/2022.wanlp-1.42.

Skiredj, Abderrahman et al. (2024). *DarijaBanking: A New Resource for Overcoming Language Barriers in Banking Intent Detection for Moroccan Arabic Speakers*. arXiv: 2405.16482 [cs.CL]. URL: https://arxiv.org/abs/2405.16482.

Tiedemann, Jörg (Nov. 2020). "The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT." In: *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, pp. 1174–1182. URL: https://aclanthology.org/2020.wmt-1.139.

Tiedemann, Jörg and Santhosh Thottingal (Nov. 2020). "OPUS-MT – Building open translation services for the World." In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61.